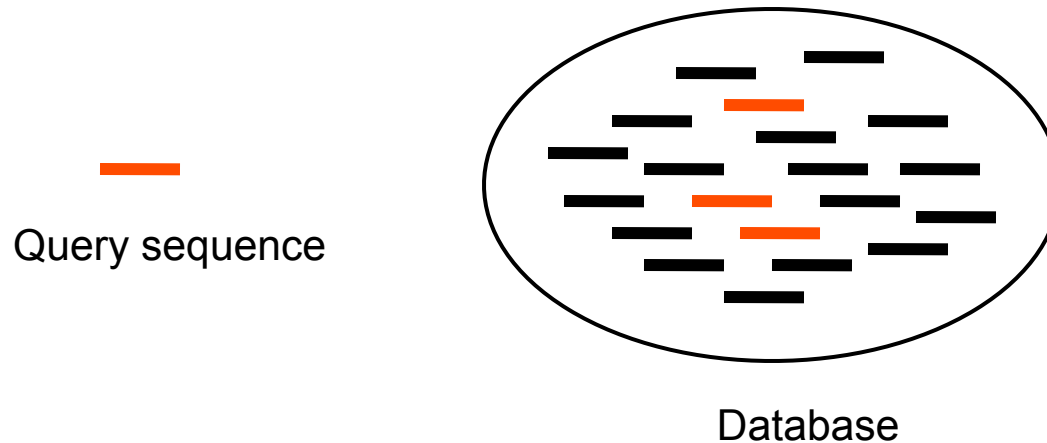

BLAST

Anders Gorm Pedersen

Database searching

**Using pairwise alignments to search
databases for similar sequences**



Database searching

Most common use of pairwise sequence alignments is to search databases for related sequences. For instance: find probable function of newly isolated protein by identifying similar proteins with known function.

Most often, ***local*** alignment (“Smith-Waterman”) is used for database searching: you are interested in finding out if ANY domain in your protein looks like something that is known.

Often, full Smith-Waterman is too time-consuming for searching large databases, so heuristic methods are used (fasta, BLAST).

BLAST: about 100 x faster than full Smith-Waterman

BLAST flavors

BLASTN

Nucleotide query sequence

Nucleotide database

BLASTP

Protein query sequence

Protein database

BLASTX

Nucleotide query sequence

Protein database

Compares all six reading frames
with the database

TBLASTN

Protein query sequence

Nucleotide database

"On the fly" six frame translation of
database

TBLASTX

Nucleotide query sequence

Nucleotide database

Compares all reading frames of
query with all reading frames of
the database

Searching on the web: BLAST at NCBI

Very fast computer dedicated to running BLAST searches

Many databases that are always up to date (e.g. NR and Human Genome)

Nice simple web interface

The screenshot displays the NCBI BLAST web interface in a browser window. The address bar shows the URL: <http://www.ncbi.nlm.nih.gov/blast/Blast.cgi?PAGE=Proteins&PROGRAM=blastp&...>. The page title is "Protein BLAST: search protein databases using a protein query". The interface includes a navigation bar with links: Home, Recent Results, Saved Strategies, and Help. A "My NCBI" link is also present. The main content area is titled "Enter Query Sequence" and contains several input fields: "Enter accession number, gi, or FASTA sequence" (with a "Clear" button), "Query subrange" (with "From" and "To" fields), "Or, upload file" (with a "Choose File" button and "no file selected" text), and "Job Title" (with a text input field). Below these is a "Choose Search Set" section with "Database" (set to "Non-redundant protein sequences (nr)"), "Organism" (with a text input field and a note about 20 top taxa), and "Entrez Query" (with a text input field). The "Program Selection" section shows "Algorithm" with radio buttons for "blastp (protein-protein BLAST)" (selected), "PSI-BLAST (Position-Specific Iterated BLAST)", and "PHI-BLAST (Pattern Hit Initiated BLAST)". A "BLAST" button is prominently displayed, followed by a checkbox for "Show results in a new window". At the bottom, there is a link for "Algorithm parameters". The footer contains copyright information and links for NCBI, NLM, NIH, and DHHS.

Why is BLAST faster? (Protein search example)

- Breaks up query sequence into smaller “words” of length 3 aa
- For each word: Determine “neighborhood” = list of other words (of same length) that have a high alignment score to starting word

Query Word ($W = 3$)

TLSHAWRLSNETDKRPFIEAERL**RDQ**HKKDYPEYKYQPRRRKNGKPGSSSEADAHSE

Determine neighborhood

RDQ 16	QDQ 12	EDQ 11	RDN 11	RDB 11	BDQ 10	RDP 10
RBQ 14	REQ 12	HDQ 11	RDD 11	ADQ 10	XDQ 10	RDT 10
RDZ 14	RDR 12	ZDQ 11	RDH 11	MDQ 10	RQQ 10	RDY 10
KDQ 13	RDK 12	RNQ 11	RDM 11	SDQ 10	RSQ 10	RDX 10
RDE 13	NDQ 11	RZQ 11	RDS 11	TDQ 10	RDA 10	DDQ 9 ...

Why is BLAST faster? (Protein search example)

- For each word in neighborhood: Find exact matches in database sequences
- If two words on same “alignment diagonal” and within 40aa: try to extend alignment in either direction to create optimal local alignment
- Various heuristics for deciding when to stop extension (stop when score drops more than X , trim back to previous maximum)

Figure 5-4. Isolated and clustered words

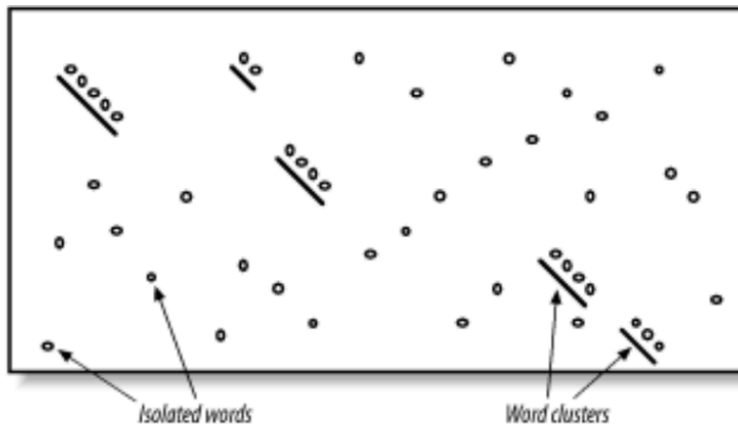
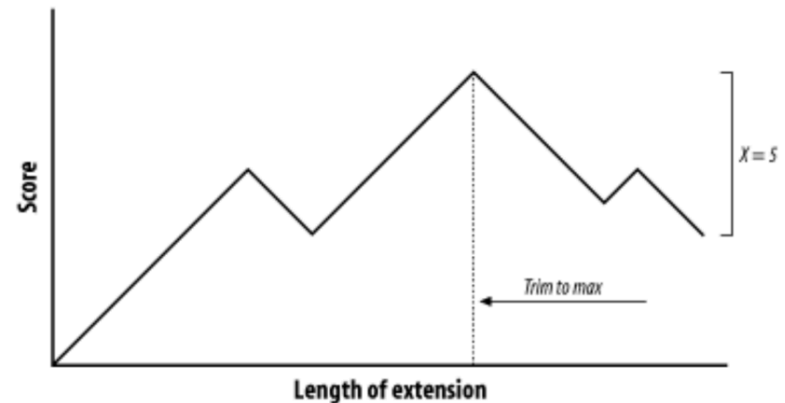


Figure 5-6. Attenuating extension with X



When is a database hit meaningful?

- **Problem:**

- Even unrelated sequences can be aligned (yielding a low score)
- How do we know if a database hit is meaningful?
- When is an alignment score sufficiently high?

- **Solution:**

- Determine the range of alignment scores you would expect to get for random reasons (i.e., when aligning unrelated sequences).
- Compare actual scores to the distribution of random scores.
- Is the real score much higher than you'd expect by chance?

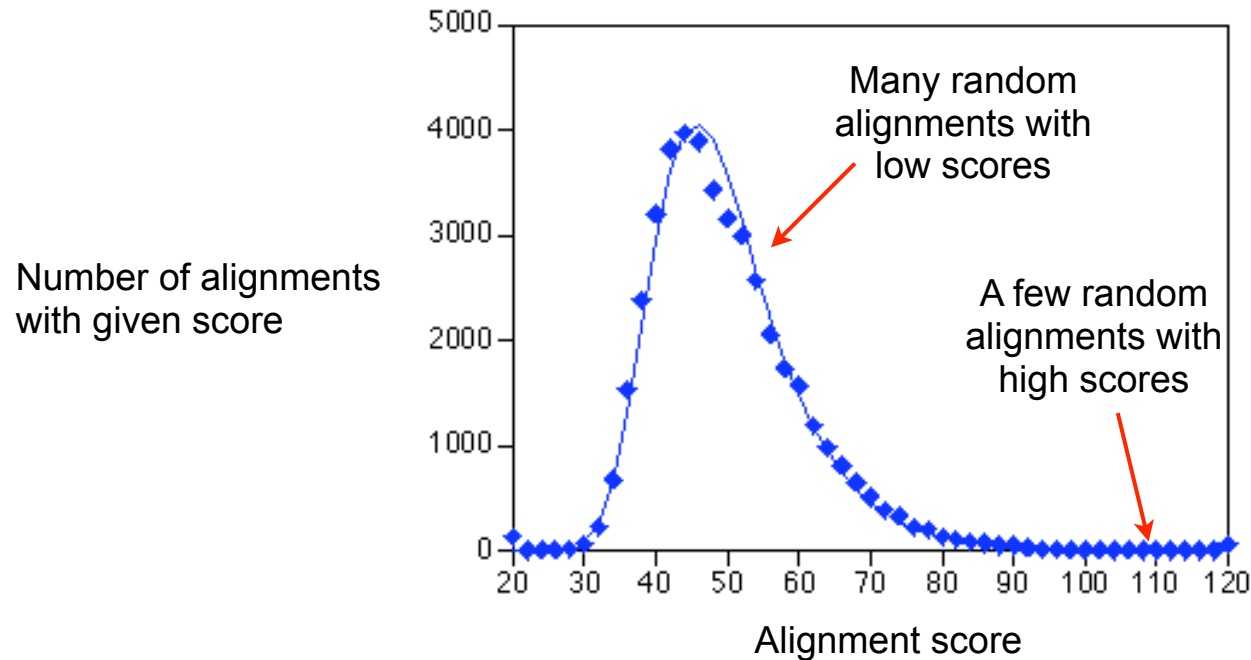
Distribution of random alignment scores

- Software simulation

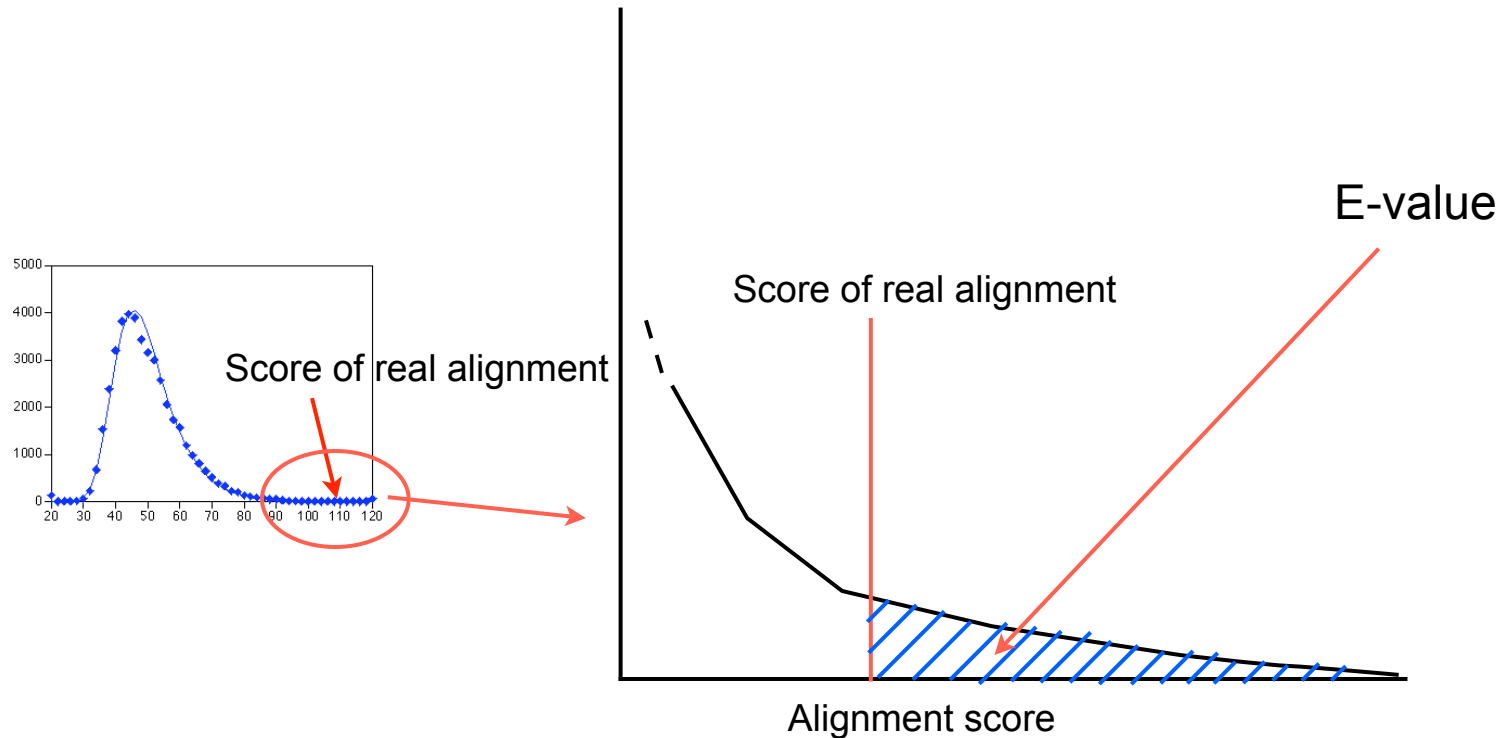
Significance of alignment score expressed as E-value

Pairwise alignment of unrelated sequences results in scores following an extreme value distribution

The exact shape and location of the distribution depends on the length and composition of the sequences



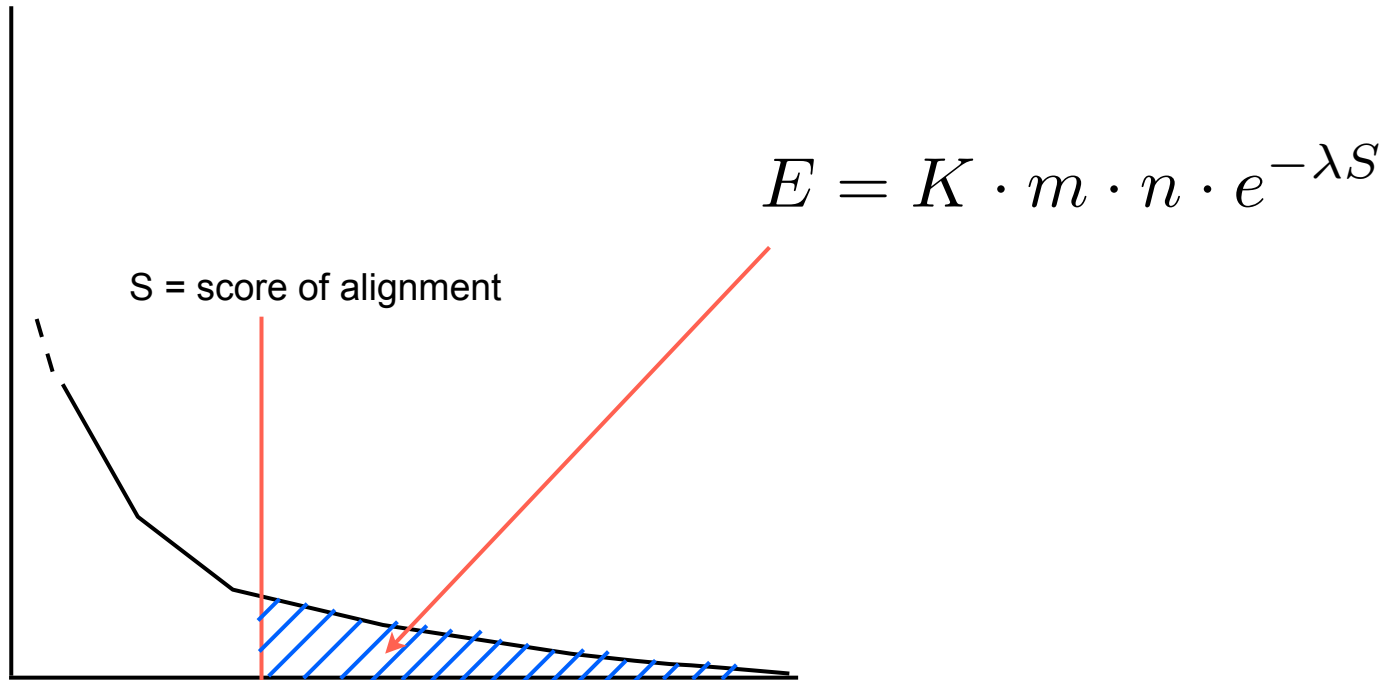
Significance of alignment score expressed as E-value



E-value: the number of random hits with score \geq real score

Want E-values well below 1 (the lower the better)

Significance of alignment score expressed as E-value



m : length of query sequence

n : combined length of all database sequences

λ : the scaling factor we also saw when computing BLOSUM62

K : a constant whose value depends on the nature of the sequences - it can be determined empirically by curve fitting